CommonWhy: A Dataset for Evaluating Entity-Based Causal Commonsense Reasoning in Large Language Models

Armin Toroghi*
University of Toronto
Toronto, ON, Canada
armin.toroghi@mail.utoronto.ca

Faeze Moradi Kalarde*
University of Toronto
Toronto, ON, Canada
faeze.moradi@mail.utoronto.ca

Scott Sanner
University of Toronto
Toronto, ON, Canada
ssanner@mie.utoronto.ca

Abstract

To effectively interact with the real world, Large Language Models (LLMs) require entity-based commonsense reasoning, a challenging task that necessitates integrating factual knowledge about specific entities with commonsense inference. Existing datasets for evaluating LLM entity-based commonsense reasoning have largely focused on True/False or multiple-choice questions, leaving the explicit assessment of the model's ability in abductive reasoning about causes and effects and generating explanations largely unexamined. In this work, we introduce CommonWhy, a dataset of 15,000 *why* questions designed to evaluate entity-based commonsense reasoning about causal relationships in LLMs. CommonWhy also serves as a Knowledge Graph Question Answering (KGQA) benchmark, as all supporting knowledge required to answer its queries is available in the Wikidata knowledge graph. Unlike existing KGQA datasets, which primarily test fact retrieval, CommonWhy targets causal commonsense reasoning, establishing a new paradigm for KGQA evaluation. Experiments with state-of-the-art LLMs and LLM-based KGQA methods reveal their significant shortcomings, including frequent factual hallucinations and failures in causal reasoning.

Keywords

Large Language Models, Commonsense Reasoning, Knowledge Graphs, Question Answering

ACM Reference Format:
Armin Toroghi, Faeze Moradi Kalarde, and Scott Sanner. 2026. CommonWhy: A Dataset for Evaluating Entity-Based Causal Commonsense Reasoning in Large Language Models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3805712.3808628

1 Introduction

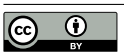
Large Language Models (LLMs) have ushered in a new phase of artificial intelligence, becoming increasingly useful as general-purpose assistants. Their effectiveness largely stems from their ability to encode vast amounts of factual knowledge [19, 23], as well as their mastery of commonsense knowledge, the general knowledge that humans develop about the world and how it works [51, 52, 64]. Entity-based commonsense reasoning [14, 36, 53], which involves commonsense reasoning about concrete entities such as people, places, and objects, captures both of these capabilities and is crucial for interacting with the real world. Therefore, evaluating the performance of LLMs on entity-based commonsense reasoning is essential.

To facilitate systematic assessment of entity-based commonsense reasoning, multiple benchmark datasets have been introduced [14, 36, 53]. Existing datasets consist of either Yes/No questions, such as *"Did Aristotle use a laptop?"*, or require the models to verify commonsense claims such as *"Harry Potter can teach classes on how to fly on a broomstick."* by giving True/False answers, which offer only a constrained measure of the model's reasoning performance. However, none of the existing datasets evaluate LLMs' ability to perform entity-based commonsense reasoning about causes and effects through *why* questions, e.g., *"Why wouldn't Katarina Barley require a visa to watch the 2024 Champions League Final in person?"*. Studying *why* questions is important because it reveals whether the model can genuinely reason about causal relationships and constraints, and focuses on explicitly evaluating the model's *thought* processes rather than a mere final answer which may be provided by simple pattern matching or random guesses.

In this paper, we introduce CommonWhy, the first dataset designed to evaluate the *abductive* (explanation-driven) causal reasoning capabilities of LLMs for entity-based commonsense reasoning, comprising 15,000 carefully curated *why* questions each associated with one or more answers. We find that CommonWhy is challenging even for the strong modern LLMs, including recent large reasoning models such as OpenAI-o3 and DeepSeek-V3.2, with models frequently failing to produce correct answers and exhibiting a high rate of hallucinations and reasoning errors. This finding is particularly striking given that these models have consistently demonstrated strong performance on existing commonsense reasoning datasets, which predominantly emphasize *deductive* or verification-style reasoning [10, 13, 18, 58]. Our results suggest that success on such benchmarks does not transfer to abductive causal reasoning in an entity-based setting, indicating that commonsense reasoning remains an unsolved challenge for LLMs.

CommonWhy spans a broad spectrum of commonsense reasoning skills, from domain-independent capabilities such as temporal reasoning and location comparison to domain-dependent reasoning about history, professions, sports, and related real-world contexts, as shown in Figure 2. Furthermore, CommonWhy targets entities exclusively from the Wikidata knowledge graph (KG) [57], and the knowledge required to answer each query is ensured to be available within Wikidata. This design enables CommonWhy to function as a benchmark for the task of answering natural language queries over KGs, known as Knowledge Graph Question Answering

*Both authors contributed equally to this research.



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808628

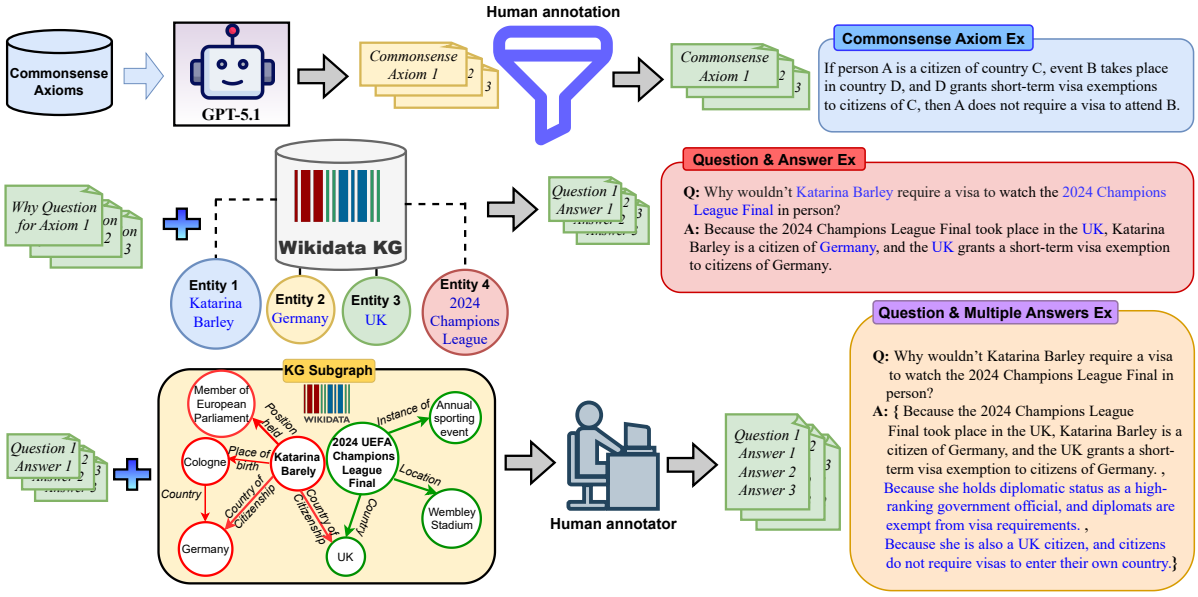


Figure 1: Overview of the CommonWhy dataset construction pipeline. (1) Commonsense axioms extracted from existing datasets are fed to GPT-5.1 to produce additional similar commonsense axioms, after which human annotators filter out invalid ones. (2) Commonsense axioms are rewritten as lifted question–answer pairs, and entities extracted from Wikidata are substituted into their variables to generate grounded pairs. (3) Human annotators are provided with the generated pairs and the corresponding KG subgraphs, and are asked to identify additional plausible causes beyond the originally generated answers.

(KGQA) [3, 55, 61, 65]. Early KGQA datasets primarily consisted of simple questions answerable using a single KG triple [61, 63]. More recent work has shifted toward multi-hop queries requiring compositional reasoning [15, 56]. However, these datasets largely focus on factoid questions that can be resolved through iterative fact retrieval and entity resolution. CoLoTa [53] partially departs from this trend by incorporating commonsense, but only in the form of deductive True/False queries. CommonWhy opens a new direction for KGQA by introducing explanation-driven abductive reasoning over KG facts that requires commonsense knowledge.

In summary, CommonWhy not only (i) evaluates LLMs on a type of commonsense reasoning that has previously been underexplored—causal, abductive, entity-based *why* reasoning—but also (ii) opens a new avenue for KGQA research in the LLM era focused on explanation-driven reasoning over KGs, inspiring future research to develop methods capable of integrating factual knowledge with commonsense inference.

2 Related Works

In this section, we provide an overview of the commonsense reasoning and KGQA datasets. Table 1 presents exemplar queries, key properties, and a comparison with CommonWhy.

2.1 Commonsense Reasoning Datasets

Humans naturally develop an understanding of how the world works, called *commonsense knowledge*, and the ability to reason with it, known as *commonsense reasoning* [31]. To interact with the real world, AI agents must also possess this knowledge and reasoning ability [2, 11, 31, 35, 54]. To evaluate AI systems' commonsense reasoning, several datasets have been developed. These datasets are divided into two categories: (i) concept-based and (ii) entity-based, which we summarize below.

(i) Concept-based Datasets evaluate commonsense reasoning about general concepts or hypothetical scenarios. CommonsenseQA [48] is a multiple-choice benchmark in which crowd workers generate questions designed to distinguish target entities associated with a source concept from ConceptNet [47]. ATOMIC [44] provides an atlas of causal commonsense rules in *if-then* form. Some of these datasets also target specific reasoning skills: PIQA [6] focuses on physical commonsense by asking which of two actions plausibly achieves a given goal, while SocialIQA [45] explores social commonsense, querying motivations, consequences, and emotional reactions in hypothetical social scenarios. Another group of works in this paradigm, such as COPA [43] and its variants [8, 40], target commonsense causal reasoning by presenting a premise with two candidate alternatives, asking the model to select the more plausible cause or effect.

(ii) Entity-based Datasets evaluate grounded commonsense reasoning about specific objects, events, or people, requiring models to combine commonsense with factual knowledge. StrategyQA [14] consists of Yes/No questions where reasoning steps are implicit and must be inferred using a strategy, with each query annotated with evidence from Wikipedia. Crowd workers generate questions given an entity term, and adversarial models are used to iteratively create harder queries. CREAK [36] similarly contains claims

Table 1: Comparison of existing commonsense reasoning and KGQA datasets with CommonWhy. CommonWhy is the first dataset that focuses on abductive causal commonsense reasoning and supports the evaluation of both LLMs and KGQA methods.

Dataset Name	Category	Exemplar Query	Commonsense	Factual	Usable for KGQA	Abductive	Answer Format
CommonsenseQA [48]	Concept-based Commonsense	*Where would I not want a fox?* ***Answer:*** *hen house*	✓	✗	✗	✗	Multiple-choice
PIQA [6]	Concept-based Commonsense	*How do I find something I lost on the carpet?* ***Answer:*** *Put a hair net on the end of your vacuum and turn it on.*	✓	✗	✗	✗	Multiple-choice
StrategyQA [14]	Entity-based Commonsense	*Did Aristotle use a laptop?* ***Answer:*** *No.*	✓	✓	✗	✗	True/False
CREAK [36]	Entity-based Commonsense	*All nuns act in holy ways.* ***Answer:*** *True.*	✓	✓	✗	✗	True/False
WebQuestions [4]	KGQA	*What country is the Grand Bahama Island in?* ***Answer:*** *Bahamas*	✗	✓	✓	✗	Entity Retrieval
LC-QuAD [56]	KGQA	*In which state is the Channel district?* ***Answer:*** *Florida*	✗	✓	✓	✗	Entity Retrieval
WikiWhy [20]	General QA	*Why did thousands of people evacuate from Northeastern China?* ***Answer:*** *The flooding triggered by Typhoon Bovalen.*	✗[1]	✓	✗	✓	Explanation (Single-answer)
CommonWhy (Ours)	Entity-based Commonsense, KGQA	*Why couldn't An Chang-ok have voted for Choi Kyu-hah to become president of South Korea?* ***Answer:*** *{"An Chang-ok is not a citizen of South Korea.", "An Chang-ok was born in 2003, long after Choi Kyu-hah became president in 1979.", "Choi Kyu-hah did not become president through a public election."}*	✓	✓	✓	✓	Explanation (Multi-answer)

about entities labeled True or False, also written by crowd workers using Wikipedia, but without a model-in-the-loop. While this human-based approach produces diverse and challenging questions, it tends to focus on popular entities, which are well-represented in the training data of LLMs. CoLoTa [53] replaces these popular entities in StrategyQA and CREAK with obscure ones to study the influence of entity popularity on the performance.

Despite the diversity of existing commonsense reasoning datasets, none of them evaluate abductive causal reasoning over entities in a way that requires explicit explanation; CommonWhy fills this gap by introducing *why* questions to entity-based commonsense reasoning. The closest existing dataset to CommonWhy is WikiWhy [20], which focuses on answering *why* questions but does not target commonsense reasoning as required for real-world decision-making. Most WikiWhy questions can be answered by extracting or paraphrasing explanations explicitly stated in Wikipedia articles, raising the risk that LLMs rely on memorization rather than genuine reasoning [5, 59]. In contrast, CommonWhy targets entity-based commonsense reasoning in settings where the relevant reasoning is not explicitly stated, requiring models to combine factual knowledge with implicit commonsense assumptions.

2.2 KGQA Datasets

KGs [21] are widely used to represent relational information about entities, with applications ranging from healthcare [41] to recommendation systems [42, 49, 50]. Retrieving information from KGs has traditionally relied on query languages such as RQL [24] and SPARQL [46], which require technical expertise and limit accessibility [52]. To address this, KGQA has emerged as a means of answering natural language queries over KGs [3, 17, 55, 61, 65].

Early KGQA datasets focused on simple questions answerable with a single KG triple, while later benchmarks introduced multi-hop reasoning over larger graphs. Representative datasets include WebQuestions [4] and its SPARQL extension, WebQuestionsSP [60] and LC-QuAD [56] grounded in DBpedia [26], and GrailQA [15] based on Freebase [7]. Despite increased structural complexity, these datasets are predominantly constructed from executable logical forms and therefore focus on factoid questions, limiting their ability to evaluate commonsense reasoning beyond iterative fact retrieval. Although CoLoTa [53] introduces commonsense into KGQA, it remains limited to deductive True/False reasoning. CommonWhy advances the task by evaluating abductive causal reasoning through explanation-based answers to natural language *why* questions.

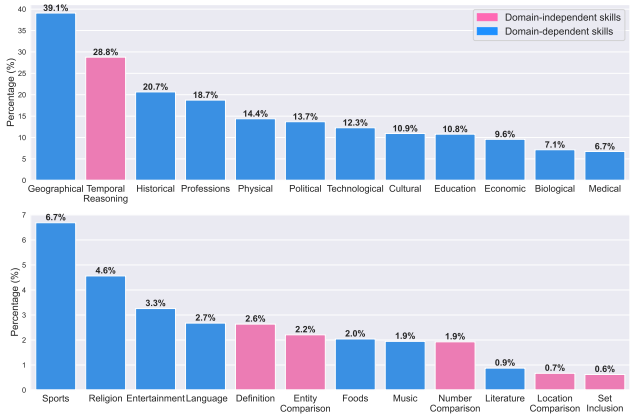


Figure 2: Distribution of reasoning skills in CommonWhy.

3 CommonWhy Dataset

In this section, we first describe the structure and properties of the CommonWhy dataset [2], and next present the methodology used for its construction.

3.1 Description

CommonWhy comprises 15,000 questions each associated with one or more answers centered on entities from the Wikidata KG, designed to evaluate abductive causal commonsense reasoning.

[1] Although some queries require commonsense reasoning, WikiWhy questions are extracted from Wikipedia articles and most queries can be answered verbatim.

[2] The dataset and code are available here.

Queries of CommonWhy cover a diverse range of commonsense reasoning skills as shown in Figure 2. Each dataset entry consists of: (i) a *why* question that targets at least one Wikidata entity; (ii) the unique Wikidata identifier (QID) of the anchor entities mentioned in the query; (iii) a relevant Wikidata subgraph containing the factual information required to answer the question; (iv) a general commonsense axiom, expressed in natural language, that captures the implicit knowledge underlying the causal relationship between the question and its answers; and (v) one or more ground-truth answers that specify possible causes for the effect described in the query. We describe each constituent in detail below.

(i) Query. A CommonWhy query is an interrogative sentence beginning with *why* that asks for the cause of an effect described in the question. Each query q is associated with $L \geq 1$ ground-truth answers $\{a_q^{(l)}\}_{l=1}^{L}$, which can be derived by combining knowledge graph facts with abductive causal commonsense reasoning.

(ii) Wikidata entities. Each query refers to a set of anchor entities drawn from Wikidata, for which factual information is required to answer the question. For each anchor entity, we provide both its label and QID, which can be used to retrieve the required facts.

(iii) Relevant KG Subgraph. Queries in CommonWhy are curated to ensure that all factual knowledge required to answer them is contained within the Wikidata KG. For each query, we provide the set of relevant facts extracted from a Wikidata subgraph, represented as triples of the form (h, r, t), in which h, r, and t denote the head entity, relation, and tail entity, respectively. These triples collectively supply the factual evidence needed to support the answer. For triples containing qualifiers in the format *relation-entity* pairs that provide additional context to the original KG triple, we represent them as hyper-relational triples $\big((h, r, t), \{(k_i, v_i)\}_{i=1}^{m}\big)$, where each (k_i, v_i) corresponds to a qualifier relation and its associated entity.

(iv) Commonsense Axiom. The commonsense axiom is a logical statement, expressed in natural language, that formally captures the causal commonsense knowledge that links an observed effect stated in the query to its plausible underlying causes. Specifically, it formalizes the properties of the entities mentioned in the query and the relations among them as premises that, when satisfied, constitute a sufficient causal explanation for the effect.

Formally, let $\mathcal{E}_q = \{e_{1,q}, \ldots, e_{|\mathcal{E}_q|,q}\}$ denote the set of entities referenced in a query q. The commonsense axiom I_q is a natural-language realization of a First-Order Logic (FOL) expression

$$a_q \wedge \left(\bigwedge_{i=1}^{|\mathcal{P}|} \bigwedge_{j=1}^{|\mathcal{E}_q|} P_i(e_{j,q}) \right) \wedge \left(\bigwedge_{i=1}^{|\mathcal{F}|} \bigwedge_{j=1}^{|\mathcal{E}_q|} F_i(e_{j,q}) \, \langle op_j^i \rangle \, e_{j,q}^i \right) \Rightarrow \; effect_q,$$

where $\mathcal{P} = \{P_1, \ldots, P_{|\mathcal{P}|}\}$ is a set of predicates, $\mathcal{F} = \{F_1, \ldots, F_{|\mathcal{F}|}\}$ is a set of functions, $\langle op_j^i \rangle \in \{=, \neq, <, \leq, >, \geq\}$ denotes a (dis)equality or comparison operator, and a_q represents a candidate cause.

Under this formulation, answering a query amounts to abductively identifying one or more causes a_q such that, together with the factual knowledge grounded in the KG, the commonsense axiom implies the effect described in the query.

(v) Answers. In CommonWhy, answers correspond to plausible causes that explain the effect described in a query. Since real-world effects may admit multiple valid causal explanations, the dataset adopts a multi-answer formulation in which each query q is associated with a set of correct answers $\{a_q^{(l)}\}_{l=1}^{L}$, with $L \geq 1$. Each answer represents a distinct abductive hypothesis that, when combined with factual knowledge from the KG and relevant commonsense assumptions, is sufficient to entail the observed effect. For example, the query *"Why wouldn't Katarina Barley need a visa to travel to the UK?"* admits multiple valid causes, such as: *because she is a German citizen and German citizens do not require visas to travel to the UK*; *because she holds diplomatic status as a high-ranking government official and diplomats are exempt from visa requirements*; or *because she is also a UK citizen and citizens do not require visas to enter their own country*. Rather than enforcing a single explanation, CommonWhy explicitly models these alternative causes as separate but equally valid ground-truth answers, reflecting the inherently non-unique and abductive nature of commonsense causal reasoning.

3.2 Methodology

The overall framework for constructing the CommonWhy dataset consists of three steps and is illustrated in Figure 1. Below, we describe each step in detail.

3.2.1 Axiom Generation. Constructing why questions that genuinely require commonsense reasoning is challenging. The only existing large-scale dataset of why questions widely used in LLM evaluation, WikiWhy [20], generates queries by extracting sentences from Wikipedia that contain explicit causal markers such as *because* or *due to*. While scalable, this approach produces questions that are often answerable by directly retrieving or paraphrasing the source text, and therefore primarily evaluates retrieval rather than implicit commonsense reasoning. Since CommonWhy targets abductive causal reasoning grounded in unstated assumptions, this extraction-based strategy is insufficient. Moreover, building commonsense reasoning datasets on a scale has traditionally required substantial human effort [22, 32]. However, recent works show that LLMs have become increasingly effective in learning and generalizing lifted commonsense axioms, with near-ideal performance in existing concept-based benchmarks [10, 18, 58]. This motivates us to leverage LLMs' capability to generate lifted commonsense axioms. For entity-grounded query generation introduced in the subsequent step, however, we rely on a KG rather than LLMs, as LLMs are prone to hallucinations and factual inaccuracies.

Concretely, we provide GPT-5.1 with a set of commonsense axioms manually extracted from the StrategyQA [14], CREAK [36], and CoLoTa [53] datasets, and use them as few-shot examples. The model is then prompted to creatively expand this set by proposing additional axioms that follow similar causal structures. For example, given an exemplar axiom such as *"If person A is from country C, and museum B is also in country C, then A does not require a visa to visit B"*, the model generates related axioms like *"If person A is a citizen of country C, event B takes place in country D, and D grants short-term visa exemptions to citizens of C, then A does not require a visa to attend B"*. To ensure quality and correctness, all LLM-generated axioms are independently reviewed by two human annotators, who verify whether each rule aligns with human commonsense. We retain only axioms unanimously judged to be valid.

3.2.2 Query Generation. After generating a diverse set of valid commonsense axioms, each axiom is rewritten into a lifted question–answer pair. Next, we ground these question-answer pairs by

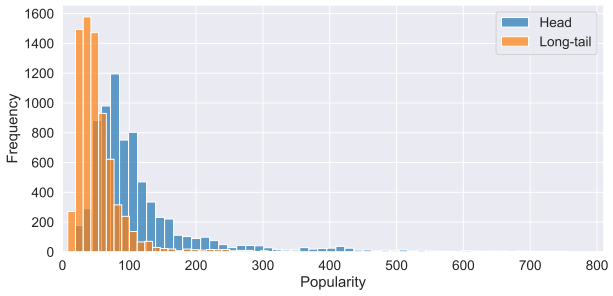


Figure 3: Entity popularities in the head and long-tail splits.

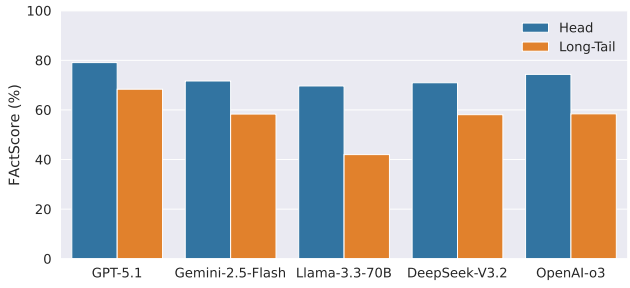


Figure 4: FActScores obtained by different LLMs across head and long-tail splits.

instantiating their abstract variables with concrete entities from Wikidata. Specifically, for each lifted pair, we identify entities that satisfy the specified properties and relations. For example, given the lifted pair *"Why wouldn't Person A require a visa to attend event B?"* with the answer *"Because event B takes place in country D, Person A is a citizen of country C, and D grants short-term visa exemptions to citizens of C"*, we must instantiate the variables A, B, C, and D with real-world entities that fulfill these constraints. To this end, we construct SPARQL [46] queries over Wikidata to retrieve candidate persons, events, and countries that jointly satisfy the required factual conditions, and extract the relevant subgraph containing the supporting triples. This grounding process results in fully instantiated, entity-specific queries paired with at least one valid answer that can only be obtained by combining factual knowledge from the KG with abductive causal commonsense reasoning.

The work in [29] has shown that questions in many QA datasets are often phrased unnaturally, and that model performance degrades significantly when evaluated on more natural reformulations. To avoid this issue, we follow their methodology and explicitly rewrite the question answers into fluent, human-like why questions, ensuring that CommonWhy evaluates reasoning ability rather than robustness to unnatural question phrasing.

3.2.3 Multi-Answer Annotation. Once queries are instantiated and grounded in Wikidata, we complete the set of plausible answers by identifying all valid causes that could explain the effect stated in the query, given the retrieved KG subgraph. Importantly, this set of causes is entity-dependent: while a lifted commonsense axiom specifies a general causal pattern, the concrete properties of the instantiated entities may introduce additional valid answers. For example, although a generic German citizen may not require a visa to attend the *2024 Champions League Final* in London solely due to their nationality, a specific individual such as *Katarina Barley* admits further valid causes, like holding diplomatic status or possessing UK citizenship. To capture this variability, two human annotators independently inspect the KG subgraphs associated with each instantiated query and enumerate all plausible causal explanations supported by the available facts and commonsense assumptions. Only those causes identified by both annotators are retained, ensuring high precision and consistency. This process yields a set of answers that reflects real-world causal reasoning, where outcomes often admit multiple, simultaneously valid explanations.

4 Experiments

4.1 Baselines

CommonWhy is designed to function as both an entity-based commonsense reasoning dataset for evaluating LLMs, as well as a KGQA dataset. Hence, we evaluate it on both LLMs and KGQA methods.

LLM Baselines. We evaluate widely-used and modern LLMs: GPT-5.1 [38], Gemini-2.5-Flash [9], Llama-3.3-70B [33], and two prominent reasoning models: OpenAI-o3 [39], and DeepSeek-V3.2 [12, 30].

KGQA Baselines. Most existing KGQA methods focus on answering factoid queries by translating them into structured languages such as SPARQL [46], and do not explicitly support commonsense reasoning. Hence, we evaluate CommonWhy using two strong LLM-based KGQA baselines that rely on LLMs to supply the required commonsense knowledge: (i) KB-Binder [27], which employs in-context learning to translate natural language queries into SPARQL, and (ii) KGR [16], which generates candidate claims with an LLM and retrofits them against the KG to enforce factual consistency. We use GPT-5.1 as the LLM backbone for both methods.

4.2 Evaluation Metrics

Answering *why* questions in CommonWhy involves generating explanations that identify plausible causes of an observed effect, which differs fundamentally from the accuracy-based evaluation used in True/False or multiple-choice commonsense benchmarks. As a result, we adopt an evaluation scheme tailored to explanation-driven causal reasoning rather than discrete answer selection.

Answer Correctness. We evaluate whether a model's response is logically equivalent to at least one of the ground-truth causal explanations provided for a query, regardless of surface-level phrasing differences. Because exhaustive human evaluation is costly, we use GPT-4o [37] as an LLM judge to determine logical equivalence between generated answers and any of the ground-truth answers.

Standard Long-form Generation Metrics. We report standard metrics for long-form generation: BERTScore (BS) [62] precision, recall, and F1; ROUGE-L [28]; and METEOR [1]. While these metrics do not directly evaluate causal reasoning, they provide complementary signals about explanation quality and help characterize how closely model outputs align with ground-truth answers. In our multi-label setting, where each query may have multiple valid ground-truth answers, we compute each metric between the model

Table 2: Performance comparison on Long-Tail and Head subsets of CommonWhy across different LLMs and KGQA methods.

Model	Long-Tail							Head						
	BS-Prec	BS-Rec	BS-F1	ROUGE-L	BLEU (%)	METEOR	Correctness (%)	BS-Prec	BS-Rec	BS-F1	ROUGE-L	BLEU (%)	METEOR	Correctness (%)
GPT-5.1 [38]	0.072	0.424	0.242	0.223	4.92	0.327	57.09	0.063	0.426	0.239	0.219	4.56	0.319	60.49
Gemini-2.5-Flash [9]	**0.277**	**0.476**	**0.375**	**0.318**	**9.23**	**0.359**	43.83	**0.274**	**0.471**	**0.371**	**0.315**	**9.13**	**0.358**	46.21
Llama 3.3-70B [33]	0.214	0.435	0.323	0.257	7.26	0.327	39.20	0.221	0.437	0.328	0.258	7.03	0.324	43.43
DeepSeek-V3.2 [12]	0.032	0.394	0.207	0.188	3.72	0.274	40.89	0.026	0.397	0.205	0.185	3.62	0.273	41.60
OpenAI-o3 [39]	0.172	0.412	0.289	0.258	6.11	0.309	**67.93**	0.161	0.417	0.286	0.254	5.78	0.309	**68.51**
KB-Binder [27]	0.009	0.122	0.042	0.099	0.911	0.145	18.48	0.014	0.181	0.083	0.096	1.88	0.156	20.45
KGR [16]	0.092	0.403	0.231	0.230	5.02	0.289	56.22	0.083	0.448	0.251	0.220	4.89	0.323	62.33

prediction and every reference answer, and report the maximum score across references.

Answer Factuality. FActScore [34] quantifies the factual consistency of an LLM-generated answer by measuring the proportion of its atomic claims that are supported by an external knowledge source C. Atomic claims are validated against Wikidata and, when unsupported, cross-checked using Wikipedia. Given an answer a_q, consisting of the set of atomic facts A_q, the FActScore is computed as $f(a_q) = \frac{1}{|A_q|} \sum_{f \in A_q} \mathbb{I}\,(f \text{ is supported by } C)$, where $|A_q|$ denotes the cardinality of the set A_q, and $\mathbb{I}(\cdot)$ is the binary indicator function that maps a true statement to 1 and a false statement to 0.

4.3 Results

To analyze the impact of entity popularity on performance, we split CommonWhy into equally-sized head and long-tail subsets based on Wikidata triple counts, following CoLoTa [53]. The popularity distributions for both splits are shown in Figure 3. This section summarizes the main results of the experiments.

CommonWhy reveals limitations of state-of-the-art LLMs in entity-based commonsense reasoning: Results of the experiments are summarized in Table 2. Overall, answer correctness is low across all evaluated models, underscoring the difficulty of abductive causal commonsense reasoning over entities. The best-performing model, OpenAI-o3, a prominent reasoning model, achieves correctness scores of only 67.93% and 68.51% on the long-tail and head splits, respectively. DeepSeek-V3.2, another well-known reasoning model, performs substantially worse, reaching 40.89% and 41.60% correctness on the two splits. These results highlight that even state-of-the-art reasoning models struggle with entity-grounded abductive causal reasoning.

Furthermore, we observe that standard long-form generation metrics are not reliable indicators of causal correctness. For example, Gemini-2.5-Flash consistently achieves the highest scores on BERTScore, ROUGE-L, and METEOR, yet its answer correctness remains lower than that of OpenAI-o3 and GPT-5.1. This discrepancy suggests that lexical or semantic similarity to reference explanations does not necessarily reflect correct abductive reasoning, indicating that standard long-form generation metrics are insufficient and that evaluating causal correctness requires LLM-based judges.

Existing KGQA methods cannot perform causal commonsense reasoning to answer CommonWhy queries: Both KGQA approaches, KB-Binder [27] and KGR [16], perform poorly on the abductive causal commonsense reasoning task introduced in CommonWhy. In particular, KB-Binder exhibits especially weak performance, as it relies on translating questions into SPARQL queries—a formulation ill-suited for answering *why* questions that require implicit commonsense and causal inference beyond explicit triples in the knowledge graph. KGR likewise fails to significantly improve over directly prompting its backbone LLM (GPT-5.1), indicating that current KGQA pipelines remain inadequate for abductive causal reasoning, highlighting the need for future methodological advances.

LLMs and KGQA methods perform worse on abductive causal commonsense reasoning over long- tail entities: Prior works have shown that LLM performance on True/False entity-based commonsense questions declines as entity popularity decreases [53]. We observe the same trend for abductive causal commonsense reasoning: Across all LLMs and KGQA methods, performance consistently declines when moving from queries involving head entities to those targeting long-tail entities. Importantly, in constructing the head and long-tail splits of CommonWhy, we control for reasoning complexity by ensuring that both splits are derived from the same set of commonsense axioms, such that the underlying reasoning patterns remain identical and only entity popularity varies. This design makes CommonWhy particularly suitable for studying and developing methods that improve the robustness of reasoning models to variations in entity popularity.

LLMs suffer from hallucinations in causal commonsense reasoning, particularly over long-tail entities. To assess the factual reliability of model outputs beyond answer correctness, we compute FActScore for the responses judged as correct, following Kazemi et al. [25]. As shown in Figure 4, all models achieve modest FActScores, indicating that even logically correct answers often contain hallucinated facts. This issue becomes more pronounced for long-tail entities. These findings highlight that CommonWhy not only challenges models' abductive causal reasoning abilities, but also exposes their tendency to introduce factual hallucinations when integrating knowledge with commonsense inference.

5 Conclusion

We proposed CommonWhy, a new dataset for evaluating the abductive causal commonsense reasoning capability of LLMs in entity-based settings. We observed a high rate of hallucinations and reasoning errors even from leading reasoning LLMs on CommonWhy. Since the factual knowledge required to answer its queries is

grounded in Wikidata, CommonWhy also establishes a new direction for KGQA: answering *why* questions that require abductive causal reasoning rather than purely factual retrieval. Our experiments with two LLM-based KGQA methods further show the insufficiency of current methods for this setting. Overall, CommonWhy offers a challenging testbed for studying causal reasoning in LLMs, and paves the way for future reasoning-oriented KGQA research.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project).

References

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[2] Marco Baroni, Armand Joulin, Allan Jabri, German Kruszewski, Angeliki Lazaridou, Klemen Simonic, and Tomas Mikolov. 2017. CommAI: Evaluating the first steps towards a useful general AI. *arXiv preprint arXiv:1701.08954* (2017).

[3] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1533–1544.

[4] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 1533–1544. https://aclanthology.org/D13-1160/

[5] Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". In *The Twelfth International Conference on Learning Representations*.

[6] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 7432–7439.

[7] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 1247–1250.

[8] Ana Brassard, Benjamin Heinzerling, Pride Kavumba, and Kentaro Inui. 2022. COPA-SSE: Semi-structured Explanations for Commonsense Reasoning. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 3994–4000.

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[10] Joseph Cotnareanu, Didier Chetelat, Yingxue Zhang, and Mark Coates. 2026. A Balanced Neuro-Symbolic Approach for Commonsense Abductive Logic. *arXiv preprint arXiv:2601.18595* (2026).

[11] Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* 58, 9 (2015), 92–103.

[12] DeepSeek-AI. 2025. DeepSeek-V3 Technical Report. *arXiv preprint* (2025). https://arxiv.org/abs/2512.02556

[13] İbrahim Ethem Deveci and Duygu Ataman. 2025. The Ouroboros of Benchmarking: Reasoning Evaluation in an Era of Saturation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

[14] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Trans. Assoc. Comput. Linguistics* 9 (2021), 346–361. doi:10.1162/TACL_A_00370

[15] Yu Gu, Sue Kase, Michelle Vanni, Brian M. Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 3477–3488. doi:10.1145/3442381.3449992

[16] Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18126–18134.

[17] Willis Guo, Armin Toroghi, and Scott Sanner. 2024. Cr-lt-kgqa: A knowledge graph question answering dataset requiring commonsense reasoning and long-tail knowledge. *arXiv preprint arXiv:2403.01395* (2024).

[18] Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics* 13 (2025), 1507–1526.

[19] Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main Volume*. 1772–1791.

[20] Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2022. Wikiwhy: Answering and explaining cause-and-effect questions. *arXiv preprint arXiv:2210.12152* (2022).

[21] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)* 54, 4 (2021), 1–37.

[22] Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. 2023. CRoW: Benchmarking commonsense reasoning in real-world tasks. *arXiv preprint arXiv:2310.15239* (2023).

[23] Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How Large Language Models Encode Context Knowledge? A Layer-Wise Probing Study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 8235–8246.

[24] Gregory Karvounarakis, Sofia Alexaki, Vassilis Christophides, Dimitris Plexousakis, and Michel Scholl. 2002. RQL: a declarative query language for RDF. In *Proceedings of the 11th international conference on World Wide Web*. 592–603.

[25] Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2023. Lambada: Backward chaining for automated reasoning in natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6547–6568.

[26] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web* 6, 2 (2015), 167–195.

[27] Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023. Few-shot In-context Learning on Knowledge Base Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6966–6980.

[28] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[29] Trond Linjordet and Krisztian Balog. 2022. Would you ask it that way? measuring and improving question naturalness for knowledge graph question answering. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3090–3098.

[30] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556* (2025).

[31] Hugo Liu and Push Singh. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal* 22, 4 (2004), 211–226.

[32] Adyasha Maharana and Mohit Bansal. 2022. GraDA: Graph generative data augmentation for commonsense reasoning. In *Proceedings of the 29th International Conference on Computational Linguistics*. 4499–4516.

[33] Meta AI. 2024. LLaMA 3.3 70B Instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[34] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 12076–12100.

[35] Robert C Moore. 1982. *The role of logic in knowledge representation and commonsense reasoning*. SRI International. Artificial Intelligence Center.

[36] Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/5737c6ec2e0716f3d8a7a5c4e0de0d9a-Abstract-round2.html

[37] OpenAI. 2024. GPT-4o Technical Report. (2024). https://cdn.openai.com/gpt-4o-system-card.pdf

[38] OpenAI. 2025. GPT-5.1. https://openai.com/index/gpt-5-1/

[39] OpenAI. 2025. OpenAI o3. https://platform.openai.com/docs/models/o3

[40] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2362–2376.

[41] Nidhi Rastogi and Mohammed J Zaki. 2020. Personal health knowledge graphs for patients. *arXiv preprint arXiv:2004.00071* (2020).

[42] Shaina Raza, Mizanur Rahman, Safiullah Kamawal, Armin Toroghi, Ananya Raval, Farshad Navah, and Amirmohammad Kazemeini. 2024. A comprehensive review of recommender systems: Transitioning from theory to practice. *arXiv preprint arXiv:2407.13699* (2024).

[43] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning.. In *AAAI spring symposium: logical formalizations of commonsense reasoning*. 90–95.

[44] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3027–3035.

[45] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4463–4473.

[46] Andy Seaborne and Eric Prud'hommeaux. 2008. SPARQL query language for RDF. *W3C Recommendation, W3C* (2008).

[47] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder Singh and Shaul Markovitch (Eds.). AAAI Press, 4444–4451. doi:10.1609/AAAI.V31I1.11164

[48] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4149–4158. doi:10.18653/V1/N19-1421

[49] Zhenwei Tang, Griffin Floto, Armin Toroghi, Shichao Pei, Xiangliang Zhang, and Scott Sanner. 2023. LogicRec: Recommendation with Users' Logical Requirements. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2129–2133.

[50] Armin Toroghi, Griffin Floto, Zhenwei Tang, and Scott Sanner. 2023. Bayesian Knowledge-driven Critiquing with Indirect Evidence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1838–1842.

[51] Armin Toroghi, Willis Guo, Ali Pesaranghader, and Scott Sanner. 2024. Verifiable, Debuggable, and Repairable Commonsense Logical Reasoning via LLM-based Theory Resolution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 6634–6652.

[52] Armin Toroghi, Willis Guo, Mohammad Mahdi Abdollah Pour, and Scott Sanner. 2024. Right for Right Reasons: Large Language Models for Verifiable Commonsense Knowledge Graph Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 6601–6633.

[53] Armin Toroghi, Willis Guo, and Scott Sanner. 2025. CoLoTa: A Dataset for Entity-based Commonsense Reasoning over Long-Tail Knowledge. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3444–3454.

[54] Armin Toroghi, Ali Pesaranghader, Tanmana Sadhu, and Scott Sanner. 2025. Llm-based typed hyperresolution for commonsense reasoning with knowledge bases. In *The Thirteenth International Conference on Learning Representations*.

[55] Armin Toroghi and Scott Sanner. 2024. Bayesian inference with complex knowledge graph evidence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 20550–20558.

[56] Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 10588)*, Claudia d'Amato, Miriam Fernández, Valentina A. M. Tamma, Freddy Lécué, Philippe Cudré-Mauroux, Juan F. Sequeda, Christoph Lange, and Jeff Heflin (Eds.). Springer, 210–218. doi:10.1007/978-3-319-68204-4_22

[57] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[58] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems* 37 (2024), 95266–95290.

[59] Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2025. On memorization of large language models in logical reasoning. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. 2742–2785.

[60] Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics. doi:10.18653/V1/P16-2033

[61] Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 201–206.

[62] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

[63] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[64] Zirui Zhao, Wee Sun Lee, and David Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. *Advances in neural information processing systems* 36 (2023), 31967–31987.

[65] Weiguo Zheng, Hong Cheng, Lei Zou, Jeffrey Xu Yu, and Kangfei Zhao. 2017. Natural language question/answering: Let users talk with the knowledge graph. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 217–226.